Tijdschrift voor Medisch Onderwijs, december 2009 | Vol. 28, nr. 6, p. 253-260

De kwaliteit van beoordelingen door simulatiepatiënten in een Objective Structured Clinical Examination (OSCE): een analyse van interbeoordelaarsovereenstemming

E.A.M. Pelgrim , E.J.P.G. Denessen, A.M. Hettinga, C.T. Postma

Samenvatting

Inleiding: *In deze studie staat de kwestie centraal of simulatiepatiënten, die getraind zijn in het spelen van de rol van patiënt tijdens examens maar niet medisch geschoold zijn, in staat zijn kandidaten bij een stationsexamen te beoordelen. De context waarbinnen dit onderzoek plaatsvond was de toetsing van medische competenties van buitenlandse artsen via het zogenaamde Dutch Clinical Skills Assessment (DCSA). De hoofdvraag was: wat is de overeenstemming in de door simulatiepatiënten gegeven beoordelingen van kandidaten bij het DCSA?*

Methode: *Voor het beantwoorden van de onderzoeksvraag zijn zowel de beoordelingen gebruikt van 445 koppels van twee simulatiepatiënten als de beoordelingen van 17 koppels van een simulatiepatiënt en een arts. Om na te gaan in hoeverre er consistentie was in de beoordelingen van een koppel zijn voor anamnese, lichamelijk onderzoek en communicatieve vaardigheden Cohen's kappa-coëfficiënten (K) berekend.*

Resultaten: *Uit de resultaten blijkt dat de beoordelaarsovereenstemming op anamnesevaardigheden en lichamelijk onderzoek redelijk tot hoog is. Wat betreft de communicatieve vaardigheden bleek de overeenstemming een stuk lager te zijn. Overigens waren er geen verschillen wat betreft de consistentie van de beoordelingen tussen koppels van simulatiepatiënten en koppels van een simulatiepatiënt en een arts.*

Conclusie/Discussie: *Hoewel de resultaten gebaseerd zijn op een steekproef van beperkte omvang, wijzen de resultaten van dit onderzoek erop dat beoordelingen van medische competenties op een efficiënte wijze kunnen worden verzameld door simulatiepatiënten als beoordelaars in te schakelen. (Pelgrim EAM, Denessen EJPG, Hettinga AM, Postma CT. De kwaliteit van beoordelingen door simulatiepatiënten in een Objective Structured Clinical Examination (OSCE): een analyse van interbeoordelaarsovereenstemming. Tijdschrift voor Medisch Onderwijs 2009;28(6):253-260.)*

Inleiding

In deze studie staat de kwestie centraal of simulatiepatiënten, die getraind zijn in het spelen van de rol van patiënt tijdens examens maar niet medisch geschoold zijn, in staat zijn kandidaten bij een stationsexamen te beoordelen. De studie richt zich op de beoordelaarsovereenstemming. Toetsing van klinische vaardigheden door middel van een stationsexamen is arbeidsintensief; deze beoordeling wordt vaak door een medisch specialist gedaan. Als de beoordelingen uitgevoerd kunnen worden door de simulatiepatiënten die ook de casus simuleren, leidt dat tot een kostenbesparing en vereenvoudiging van de logistiek.

De context waarbinnen dit onderzoek plaatsvond was de toetsing van de medische competenties van buitenlandse artsen. Ter waarborging van de kwaliteit van de Nederlandse gezondheidszorg kunnen medici

die hun artsdiploma in landen buiten de Europese Economische Ruimte (EER) – dit is de Europese Unie plus Noorwegen, Liechtenstein, Zwitserland en IJsland – hebben behaald, niet zonder meer als arts aan de slag in Nederland. Om vast te stellen of een buitenlandse arts al dan niet een bijscholingstraject moet volgen is een assessmentprocedure ontwikkeld. Deze procedure voor artsen van buiten de EER is in december 2005 geïmplementeerd. De procedure bestaat uit meerdere deeltoetsen waarbij één gericht is op de klinische vaardigheden en het klinisch handelen: het Dutch Clinical Skills Assessment (DCSA). Deze deeltoets is gebaseerd op het klinische assessment dat de Verenigde Staten hanteren voor buitenlandse artsen, het Clinical Skills Assessment (CSA), dat onderdeel is van de assessmentprocedure van de Educational Commission for Foreign Medical Graduates (ECFMG). In de Verenigde staten is al in 1998 begonnen met het op deze manier screenen van buitenlandse artsen.[1] De procedure die in Nederland vóór 2005 werd gebruikt was minder transparant.

Het DCSA stelt evenals het CSA kandidaten in de gelegenheid te laten zien wat hun competenties zijn op verschillende onderdelen: het afnemen van een anamnese, het verrichten van lichamelijk onderzoek, de communicatieve vaardigheden, het professioneel gedrag en de schriftelijke verslaglegging van de bevindingen inclusief het opstellen van een differentiaal diagnose en een plan van aanpak.[1] De beoordeling van anamnese en lichamelijk onderzoek is medisch-inhoudelijk; bij de communicatieve vaardigheden gaat het om de vorm. Bij het DCSA doorlopen kandidaten op één dag tien verschillende stations, bemand door simulatiepatiënten. Per station krijgen de kandidaten de opdracht om in maximaal twintig minuten een anamnese af te nemen en een lichamelijk onderzoek uit te voeren. Vervolgens heeft men tien minuten tijd om de bevindingen uit te werken en een probleemlijst, een differentiaaldiagnose en een plan van aanpak op te stellen. De simulatiepatiënt gebruikt deze tien minuten om de betreffende kandidaat op een checklist te scoren wat betreft de anamnesevaardigheden en het lichamelijk onderzoek. Ook worden de communicatieve vaardigheden en het professioneel gedrag op een beoordelingsschaal ingevuld.

Alhoewel in het CSA van de EFMG simulatiepatiënten worden ingezet voor het beoordelen van de kandidaten is het in Nederland nog niet echt gangbaar. In het onderzoek van Van der Vleuten et al. uit 1989 werd het toegepast.[2] Zij constateerden een voldoende mate van beoordelaarsovereenstemming binnen koppels van twee simulatiepatiënten. Ook in de VS waren de ervaringen met het beoordelen door simulatiepatiënten positief.[1] Het is gebruikelijk om dergelijke beoordelingen te laten doen door een gespecialiseerde arts, die het door de kandidaat verrichte anamnesegesprek en het lichamelijk onderzoek observeert.[3] Beoordelingen door artsen zijn echter tijdrovend en kostbaar. Het laten beoordelen van kandidaten door simulatiepatiënten is efficiënter maar sommigen zetten vraagtekens bij deze manier van beoordelen. Huddle is één van degenen die de inzet van simulatiepatiënten bij het beoordelen van medische vaardigheden in twijfel trekken.[4] Hij is van mening dat competenties moeten worden beoordeeld door inhoudsdeskundigen daar zij de enigen zijn die een relevant perspectief delen met de kandidaten. In zijn optiek kunnen in de medische setting de beoordelingen door simulatiepatiënten wel een rol spelen, maar worden deze uit een heel andere invalshoek gegeven als het gaat om de kwaliteit van de arts en de gezondheidszorg.

Wil men met de inzet van simulatiepatiënten kwalitatief goede beoordelingen realiseren dan zal er veel moeten worden

geïnvesteerd in het trainen van deze groep. De meeste trainingstijd blijkt te zitten in het leren scoren van de kandidaten op de checklist.[5] Nieuwe simulatiepatiënten kunnen na twee à drie uur een rol spelen en naarmate er meer ervaring is met het spelen van een patiëntrol wordt de trainingstijd korter. Ondanks het feit dat er een goede training voorafgaat aan het observeren van vaardigheden, blijven instrumenten die gebruik maken van directe observatie erg vatbaar voor observatieverschillen.[6] Wanneer verschillende beoordelaars geen overeenkomstige scores geven, wordt de beoordeling als niet accuraat beschouwd.[7] Er wordt gestreefd naar een zo hoog mogelijke overeenstemming tussen de bij een examen ingezette simulatiepatiënten.

De onderzoeksvraag van dit onderzoek was: wat is de overeenstemming in de beoordelingen van kandidaten door simulatiepatiënten die ingezet werden bij het DCSA? Het doel van het onderzoek was door het meten van de beoordelaarsovereenstemming inzicht te krijgen in de kwaliteit van de beoordelingen door simulatiepatiënten. Wanneer simulatiepatiënten accuraat blijken te scoren, zou dat een aanwijzing kunnen zijn dat ze kunnen worden ingezet bij de beoordelingen. Dit biedt, zoals gezegd, logistieke en financiële voordelen, waardoor grootschaliger toetsing mogelijk wordt. Om na te gaan in hoeverre de beoordelingen van simulatiepatiënten afwijken van de beoordelingen door een arts is ook de beoordelaarsovereenstemming tussen een simulatiepatiënt en een arts onderzocht.

Methode

Voor het beantwoorden van de onderzoeksvraag zijn de kandidaten beoordeeld door een koppel, bestaand uit een simulatiepatiënt en een observator. De observator was een simulatiepatiënt die ook was opgeleid voor de betreffende rol of een arts. De simulatiepatiënt en de observator gaven onafhankelijk van elkaar een oordeel over de kandidaat.

De gegevens zijn verzameld in de periode van 24 juni 2006 tot en met 19 januari 2008. In totaal zijn de beoordelingen van 27 simulatiepatiënten en één arts op 21 verschillende stations verzameld. De in het onderzoek betrokken simulatiepatiënten waren al langer bij het UMC St. Radboud in Nijmegen als zodanig werkzaam. Ze hadden ervaring met het spelen van patiëntrollen tijdens de klinische scholing in de reguliere opleiding tot basisarts. Voor het DCSA hebben de simulatiepatiënten een extra training gehad, die met name gericht was op het gebruik van gespecificeerde checklists voor het beoordelen van kandidaten. De scorende arts was een huisarts die zowel nauw betrokken is geweest bij de ontwikkeling van de stations in het DCSA als bij de training van de simulatiepatiënten. Omdat de observatoren onafhankelijk van elkaar beoordeelden heeft de arts-observator de scores van de simulatiepatiënt niet kunnen beïnvloeden.

Er werd in het onderzoek gebruik gemaakt van een checklist voor *anamnesevaardigheden* en een checklist voor het *lichamelijk onderzoek*. Deze waren op alle 21 stations verschillend en afhankelijk van de casus varieerden de checklists voor de anamnesevaardigheden tussen de vijf en de 17 items per station waarop 'ja' (score 1) of 'nee' (score 0) kon worden gescoord. De beoordelaar kon aankruisen of een anamnesevraag wel of niet door de kandidaat werd gesteld. Er werd een 'ja' gescoord als ook het antwoord adequaat was afgewacht. De checklisten voor de vaardigheden van het verrichten van een lichamelijk onderzoek hadden tussen de 8 en 24 items per station waarop de beoordelaar kon aankruisen of een onderzoek 'adequaat' (score 1), 'inadequaat' (score 0), of 'niet' (score 0) was uitgevoerd. Aan de categorieën

Tabel 1. *Aantal koppels dat de beoordelingen heeft uitgevoerd.*

Type koppels	Anamnese	Lichamelijk onderzoek	Communicatieve vaardigheden
Twee simulatiepatiënten	445	264	346
Een simulatiepatiënt en een arts	17	15	15
Totaal	462	279	361

'inadequaat' en 'niet' werden geen punten toegekend. Alleen als de handeling adequaat was uitgevoerd werd een punt toegekend.

Voor het beoordelen van de *communicatieve vaardigheden* werd een beoordelingsschaal toegepast. Bij alle stations werd dezelfde lijst gebruikt. Deze casusonafhankelijke lijst bestond uit 18 items. Beoordelaars konden aankruisen of de kandidaat 'goed' (score 1), 'matig' (score 0) of 'slecht' (score 0) scoorde op de verschillende subitems van communicatieve vaardigheden. In de analyse zijn evenals bij het lichamelijk onderzoek de categorieën 'matig' en 'slecht' samengenomen en werd alleen een punt toegekend bij een goede uitvoering van de betreffende communicatieve vaardigheid.

In Tabel 1 staan de gegevens over het aantal koppels dat de beoordelingen gedurende de onderzoeksperiode heeft uitgevoerd. Deze beoordelingen zijn niet alleen gedaan bij buitenlandse artsen, de groep die beoordeeld werd door middel van de toets, maar ook bij een referentiegroep van Nederlandse basisartsen, die parallel aan de buitenlandse artsen het examen aflegden; deze groep vertegenwoordigt 41% van het onderzoeksmateriaal. Om na te gaan in hoeverre de koppels van beoordelaars overeenkomstige beoordelingen toekenden zijn voor anamnese, lichamelijk onderzoek en communicatieve vaardigheden Cohen's kappa-coëfficiënten (K) berekend. Cohen's kappa is een maat voor overeenstemming, gebaseerd op de proportieovereenkomende beoordelingen die zijn gecorrigeerd voor de proportiesovereenkomsten, die op basis van toeval kon worden verwacht. Conform de richtlijnen van Cohen (1988) zijn de kappa-coëfficiënten geëvalueerd. Bij een K groter dan .75 spreken we van een grote mate van overeenstemming, bij een K tussen de .40 en .75 van een redelijke tot goede overeenstemming en bij een K lager dan .40 van een zwakke overeenstemming.

Resultaten

Voor het beantwoorden van de onderzoeksvraag is aan de hand van de bepaling van de kappa-coëfficiënten van koppels van beoordelaars nagegaan in hoeverre sprake was van een beoordelaarsovereenstemming. De resultaten van de analyses van de overeenstemming van beoordelingen van koppels van twee simulatiepatiënten staan in Tabel 2. De overeenstemming is per station bepaald. Ook zijn hierin weergegeven Cohen's kappa van het station met de laagste mate van overeenstemming, Cohen's kappa van het station met de hoogste mate van overeenstemming en de gemiddelde kappa-coëfficiënt, berekend over alle stations.

Beoordelingen van de anamnese

Uit Tabel 2 blijkt dat op het station waar de kleinste mate van overeenstemming is waargenomen sprake was van een redelijke mate van overeenstemming (K = .66). Op

Tabel 2. *Beoordelaarsovereenstemming van de beoordeling van medische competenties door 445 koppels van twee simulatiepatiënten.*

	Station met de laagste betrouwbaarheid	Station met de hoogste betrouwbaarheid	Gemiddelde betrouwbaarheid over alle stations
Anamnese	.66	.91	.78
Lichamelijk onderzoek	.49	.93	.73
Communicatieve vaardigheden	.33	.83	.59

het station met de hoogste mate van overeenstemming bleek er een grote mate van overeenstemming (K = .91). Gemiddeld over alle stations bleek dat de beoordelingen van de anamnesevaardigheden van de kandidaten door koppels van simulatiepatiënten een grote mate van overeenstemming vertoonden (gemiddelde K = .78).

Beoordelingen van het lichamelijk onderzoek

Uit Tabel 2 blijkt dat op het station waar de kleinste mate van overeenstemming is waargenomen (.49). sprake was van een redelijke mate van overeenstemming. Deze is wel lager dan de mate van overeenstemming op anamnesevaardigheden. Op het station waar de hoogste mate van overeenstemming is waargenomen was K .93, deze is iets hoger dan de hoogste mate van overeenstemming op anamnesevaardigheden. Gemiddeld is er sprake van een redelijke mate van overeenstemming over alle stations op het onderdeel lichamelijk onderzoek (.73).

Beoordelingen van de communicatieve vaardigheden

Bij deze beoordelingen zijn de kappa-coëfficiënten lager dan op anamnesevaardigheden en lichamelijk onderzoek. Op het station met de laagste mate van overeenstemming is K .33, wat een lage mate van overeenstemming betekent. Op het station met de hoogste mate van overeenstemming is wel een grote mate van overeenstemming waar te nemen (.83), al is deze lager dan de waarden op anamnese en lichamelijk onderzoek. Gemiddeld over alle stations is er een redelijke mate van overeenstemming op communicatieve vaardigheden (.59).

Tabel 3 geeft een overzicht van de gemiddelde mate van overeenstemming voor de beide typen koppels afzonderlijk. De gegevens in de bovenste rij zijn de gegevens zoals ook in Tabel 2 te zien zijn in de laatste kolom. De gemiddelde kappa-coëfficiënten voor anamnesevaardigheden, lichamelijk onderzoek en communicatieve vaardigheden bij de beoordeling door koppels van twee simulatiepatiënten waren respectievelijk .78, .73 en .59. De gemiddelde kappa voor de koppels van een simulatiepatiënt en een arts was voor de anamnese .72, voor het lichamelijk onderzoek .76 en voor de communicatieve vaardigheden .56. In de onderste rij zijn de gemiddelde kappa-coëffciënten voor de twee subgroepen samen weergegeven. Uit Tabel 3 blijkt dat de verschillen tussen beide typen koppels klein zijn. Ten opzichte van de koppels van twee simulatiepatiënten waren de gemiddelde kappa-coëfficiënten voor het koppel van een simulatiepatiënt en een arts iets lager voor anamnese, hoger

Tabel 3. *Beoordelaarsovereenstemming van de beoordeling van medische competenties door 445 koppels van twee simulatiepatiënten en 17 koppels van een simulatiepatiënt en een arts, gemiddeld over alle stations.*

	Anamnese	Lichamelijk onderzoek	Communicatieve vaardigheden
Twee simulatiepatiënten	.78	.73	.59
Een simulatiepatiënt en een arts	.72	.76	.56
Totaal	.78	.73	.59

voor het lichamelijk onderzoek en lager voor communicatieve vaardigheden. Hoewel op anamnese en lichamelijk onderzoek een mate van overeenstemming tussen .70 en .80 is waar te nemen, bleven de scores op communicatieve vaardigheden bij beide typen koppels iets onder de .60 steken.

Conclusie/Discussie

Wanneer toetsing wordt gebruikt om vergaande beslissingen te nemen over toelating of bijscholing, moet worden getracht de meetfouten zo klein mogelijk te houden. Deze studie heeft inzicht gegeven in de meetfouten die betrekking hebben op de beoordeling van medische competenties door observatie. De kwaliteit van beoordelingen van kandidaten door simulatiepatiënten bij het DCSA werd onderzocht door binnen de koppels de overeenstemming in beoordeling te meten. Daarnaast werd onderzocht in hoeverre de overeenstemming in beoordeling verschilde tussen koppels van twee simulatiepatiënten en koppels van een simulatiepatiënt en een arts.

Uit de resultaten blijkt dat de beoordelaarsovereenstemming op anamnesevaardigheden en lichamelijk onderzoek redelijk tot hoog was. Ook op de stations met de laagste mate van beoordelaarsovereenstemming werd een redelijke mate van overeenstemming bereikt. Afhankelijk van het doel van de toets zal moeten worden besloten in hoeverre deze niveaus van overeenstemming aanvaardbaar zijn, en of het wenselijk is om methoden te ontwikkelen en in te zetten om de overeenstemming te vergroten.

Alhoewel de overeenstemming van beoordelingen relatief hoog was, bleek dat de overeenstemming van de beoordelingen van communicatieve vaardigheden lager was dan van de beoordelingen van anamnesevaardigheden en het lichamelijk onderzoek. Bovendien verschilden de overeenstemmingsmaten per station. De overeenstemming op het station met de laagste mate van overeenstemming was zwak. Oorzaken van de verschillen in beoordeling op dit onderdeel zouden kunnen worden veroorzaakt door de beoordelaars (zowel arts als simulatiepatiënt), de beoordelingslijst of individuele items op deze lijst. De beoordelaars zouden niet voldoende opgeleid kunnen zijn om deze vaardigheden te beoordelen. Een andere verklaring voor de slechtere overeenstemming van de beoordelingen van communicatieve vaardigheden zou kunnen worden gezocht in de moeilijkheid van het beoordelen van deze vaardigheden. Bij het scoren van anamnesevaardigheden of lichamelijk onderzoek moest worden aangegeven of een vraag wel of niet is gesteld en of een lichamelijk onderzoek adequaat, inadequaat of niet uitgevoerd is. Daarentegen moest bij de beoordeling op communicatieve vaardigheden worden aangegeven hoe goed iemand presteert op een bepaalde vaardig-

heid. De definities van de 18 items van de beoordelingslijst voor communicatieve vaardigheden zijn wel geconcretiseerd, maar door de aard van de vaardigheid niet zo eenduidig te interpreteren als de items met betrekking tot de anamnese en het lichamelijk onderzoek. De oorzaak van een lage overeenstemming zou dus kunnen liggen aan de specificiteit van de items van een beoordelingslijst of aan het type vaardigheid dat wordt beoordeeld.

Ook het gebruik van Cohen's kappa kan invloed hebben gehad op de resultaten. Hoewel kappa aantrekkelijk is omdat hij corrigeert voor overeenstemming tussen beoordelaars op basis van toeval, blijkt kappa een te negatief beeld te geven wanneer het fenomeen zeer frequent of juist zeer weinig wordt geobserveerd.[8] Deze prevalentieafhankelijkheid van kappa wordt in ons geval waargenomen bij zeer goede of juist zeer slechte kandidaten.

De uitkomsten van het onderzoek naar het DCSA liggen in lijn met de uitkomsten van het onderzoek van Van der Vleuten et al. uit 1989.[2] Ook in hun onderzoek werd de overeenstemming tussen simulatiepatiënten bij een praktische vaardigheidstoets bepaald. Zij constateerderden eveneens een voldoende mate van beoordelaarsovereenstemming bij koppels van twee simulatiepatiënten. Er werd geen onderzoek gedaan naar de overeenstemming bij koppels van simulatiepatiënten en artsen.

Een vergelijking van de overeenstemming bij koppels van twee simulatie-patiënten en koppels van een simulatie-patiënt en een arts toonde aan dat er geen groot verschil was in de beoordelaarsover-eenstemming op de onderdelen anamnese, lichamelijk onderzoek en communicatieve vaardigheden.

Deze onderzoeksresultaten geven een eerste aanzet voor het inzetten van simulatiepatiënten bij het DCSA en soortgelijke toetsing. Dit geeft financieel en logistiek veel meer mogelijkheden. Verder onderzoek, met de inzet van simulatiepatiënten en artsen op grotere schaal, is nodig om de conclusies uit bovenstaand onderzoek te ondersteunen. Daarnaast zou het interessant zijn te kijken naar de beoordelaarsovereenstemming bij koppels van artsen. Is deze in dezelfde orde van grootte als die tussen simulatiepatiënten binnen een koppel? En welke factoren lijken daarbij van belang?

Hoewel de resultaten een beeld geven over de beoordelaarsovereenstemming in het DCSA heeft het onderzoek een aantal beperkingen. Met name het aantal koppels van een simulatiepatiënt en een arts was erg gering in vergelijking met het aantal koppels van twee simulatiepatiënten. Bovendien was slechts één arts in dit onderzoek betrokken. Om stelliger uitspraken te kunnen doen over de kwaliteit van beoordelingen door simulatiepatiënten vergeleken met die van artsen, is vervolgonderzoek nodig met een grotere steekproef van artsen.

Dit onderzoek vormde een eerste aanzet tot evaluatie van het DCSA. Hoewel de resultaten gebaseerd zijn op een steekproef van beperkte omvang, wijzen de resultaten van dit onderzoek erop dat beoordelingen van medische competenties op een efficiënte wijze kunnen worden verzameld door simulatiepatiënten als beoordelaars in te schakelen.

Literatuur

1. Peitzman SJ. Clinical skills assessment using standardized patients: Perspectives from the educational commission for foreign medical graduates. Am J Phys Med Rehabil 2000;79:490-493.
2. Vleuten CPM van der & Luyk SJ van. Beoordelen van praktische vaardigheden. In: H.J.M. van Berkel & A.E. Bax (red.). Beoordelen in het onderwijs: Een handleiding voor het construeren van toetsen en het evalueren van leerdoelen en onderwijsvormen. Houten/Zaventem: Bohn Stafleu van Loghum; 1993. [Assessment in education: A manual for test construction and evaluation of learning objectives and educational methods. Houten/Zaventhem: Bohn Stafleu van Loghum; 1993].

3. LeBlanc VR, Tabak D, Kneebone R, Nestel D, Mac-Rae H & Moulton CA. Psychometric properties of an integrated assessment of technical and communication skills. Am J Surg 2009;197:96 –101.
4. Huddle TS. The limits of objective assessment of medical practice. Theor Med Bioeth 2007;28:487-496.
5. Howard S, & Barrows MD. An overview of the uses of standardized patients for teaching and evaluating clinical skills. Acad Med 1993;68:443-451.
6. Fraenkel JR, & Wallen NE. How to design and evaluate research in education. New York: McGraw-Hill; 2006.
7. Kroft G van der & Weeren J van. Competenties meetbaar maken. Vernieuwing. Tijdschrift voor Onderwijs en Opvoeding 2006;63:13-15. [Making competencies measurable. Innovation. Journal of Teaching and Education 2006;63:13-15].
8. Pols J. & Bosveld HEP. Beoordelaarsbetrouwbaarheid (niet) meten met behulp van Cohen's kappa? Tijdschrift voor Medisch Onderwijs 2003; 22(5):229-234. [(Refrain from) measuring observer agreement with Cohen's kappa? Dutch Journal of Medical Education 2003;22(5):229-234].

De auteurs:
E A.M Pelgrim, MSc is onderwijskundige, Onderwijsinstituut UMC St. Radboud Nijmegen.
Dr. E.J.P.G. Denessen is onderwijskundige, Behavioural Science Institute, Radboud Universiteit Nijmegen.
Drs. A.M. Hettinga is huisarts en wetenschappelijk docent, Onderwijsinstituut UMC St. Radboud Nijmegen.
Dr. C T. Postma is internist, afdeling Interne Geneeskunde en Onderwijsinstituut UMC St. Radboud Nijmegen.

Correpondentieadres:
E.A.M. Pelgrim, Onderwijsinstituut UMC Nijmegen, huispostnummer 166, postbus 9101, 6500 HB Nijmegen. Tel.: 024-3610291; e-mail: e.pelgrim@elg.umcn.nl

Belangenconflict: geen gemeld
Financiële ondersteuning: geen gemeld

Summary

Introduction: *Standardized patients are trained to portray the role of a patient in a standardized and consistent way. An important question is whether standardized patients are capable of scoring the medical competencies of candidates in an examination setting. In this study we assessed the reliability of the scoring of history taking, physical examination and communication skills by standardized patients in a high stakes examination of foreign medical graduates. The main question was: What is the consensus in judgment among standardized patient in the Dutch Clinical Skills Assessment (DCSA)?*
Methods: *The data used in this research were judgments of 445 dyads of two standardized patients and 17 dyads of a standardized patient and a physician. Inter-rater agreement was assessed to examine the consistency of judgments regarding history taking, physical examination and communication skills.*
Results: *The results showed that the dyads reached relatively high levels of inter-rater agreement on all three test domains, although the agreement on communication skills was lower than the consensus on history taking and physical examination. No significant differences were observed between the levels of agreement of the dyads of two standardized patients compared to the dyads of a standardized patient and a physician.*
Conclusion/Discussion: *Although the size of our sample was fairly small – especially concerning the dyads of a standardized patient and a physician - the results of this study point at the possibility of using standardized patients as raters of medical competencies in examination settings. This could contribute to the efficiency of examination procedures. (Pelgrim EAM, Denessen EJPG, Hettinga AM, Postma CT. The quality of assessment by standardized patients in an Objective Structured Clinical Examination (OSCE): an analysis of observer agreement. Dutch Journal of Medical Education 2009;28(6):253-260.)*